Psicothema

Psicothema 2019, Vol. 31, No. 4, 458-464
doi: 10.7334/psicothema2019.116

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Methodological quality checklist for studies based on observational methodology (MQCOM)

Salvador Chacón-Moscoso[1,2], M. Teresa Anguera[3], Susana Sanduvete-Chaves[1], José L. Losada[3], José A. Lozano-Lozano[2], and Mariona Portell[4]

[1] Universidad de Sevilla, [2] Universidad Autónoma de Chile (Chile), [3] Universidad de Barcelona, and [4] Universitat Autònoma de Barcelona

Abstract

Background: No existing instrument addresses the minimum number of items that guarantee methodological quality in studies based on observational methodology. Consequently, professionals who are not experts in observational methodology do not have a basic framework to guide their practice in this type of study. This study developed a checklist to measure the minimum number of items for methodological quality that studies based on observational methodology should consider and provided evidence of their validity based on test content and intercoder reliability. **Method:** Fifty-four judges with at least 1 year of experience in observational methodology and research based on this methodology evaluated the items of the developed checklist in terms of relevance, usefulness, and feasibility. Items were selected if they obtained at least .5 in the Osterlind indexes of the three aspects evaluated. Two coders applied the selected items to a random selection of articles that used observational methodology to investigate soccer, and intercoder reliability was examined using Cohen's kappa (κ) coefficients. **Results:** The final checklist included 16 items grouped into 11 criteria/dimensions, with adequate reliability coefficients. **Conclusions:** This study developed a useful instrument for non-expert professionals to enhance the methodological quality of studies based on observational methodology.

Keywords: Methodological quality, observational methodology, checklist, intercoder reliability, validity evidence based on test content.

Resumen

Checklist de calidad metodológica para estudios basados en metodología observacional. **Antecedentes:** no existen instrumentos referidos a los ítems mínimos que garanticen la calidad metodológica en estudios basados en metodología observacional. En consecuencia, los profesionales no expertos en metodología observacional no disponen de una guía básica que oriente su práctica en este tipo de estudios. Se desarrolló una escala para medir los ítems mínimos de calidad metodológica que deben considerar los estudios basados en metodología observacional aportando evidencias de su validez basada en el contenido de la prueba y fiabilidad intercodificadores. **Método:** cincuenta y cuatro jueces con al menos un año de experiencia en metodología observacional y su aplicación evaluaron los ítems del checklist elaborado respecto a su relevancia, utilidad y viabilidad. Se seleccionaron aquellos ítems que obtuvieron al menos ,5 en los índices de Osterlind en los tres aspectos evaluados. Dos codificadores los aplicaron a una selección aleatoria de artículos que utilizaron metodología observacional en fútbol y se estudió la fiabilidad intercodificadores mediante coeficientes kappa (κ) de Cohen. **Resultados:** la escala resultante constó de 16 ítems agrupados en 11 criterios/dimensiones, con coeficientes de fiabilidad adecuados. **Conclusiones:** se desarrolló un instrumento útil dirigido a profesionales no expertos para potenciar la calidad metodológica de estudios basados en metodología observacional.

Palabras clave: calidad metodológica, metodología observacional, *checklist*, fiabilidad intercodificadores, evidencia de validez basada en el contenido del test.

This paper is product of the results of meta-analyses conducted by our research group over the past 30 years, which investigates methodological quality in studies using observational methodology. Previous related results were published in 2015, 2016, and 2018, wherein we reviewed the literature about methodological quality of primary studies in meta-analysis and presented protocols for conducting and reporting studies based on observational methodology. These advances have been presented regularly in meetings organized by the European Association of Methodology (EAM), *Asociación Española de Metodología de las Ciencias del Comportamiento* (Spanish Association of Methodology in Behavioral Sciences, AEMCCO), Methodological Advances in Social Interaction (MASI), and the Society for Research Synthesis Methods (SRSM).

This paper focuses on studies based on observational methodology, which provide a systematic and quantified recording of behavior in its natural context using non-standardized instruments, and can be adapted to any context (Anguera, 1979, 1996, 2003a). Such studies are different from those referred to in literature as *observational studies*, which usually involve implementation of a treatment or intervention without random

Received: April 23, 2019 • Accepted: July 30, 2019
Corresponding author: Susana Sanduvete-Chaves
Facultad de Psicología
Universidad de Sevilla
41018 Sevilla (Spain)
e-mail: sussancha@us.es

assignment of participants to conditions (e.g., cohort studies, case studies, cross-sectional studies, etc.).

Studies based on observational methodology have made enormous progress because of the great advances in software for record and analysis of data and because of the applicability of the studies to different areas of intervention, such as the social field (Santoyo, Jonsson, Anguera, & López-López, 2017), education (Gimeno, Anguera, Berzosa, & Ramírez, 2006), sports (Garzón, Lapresa, Anguera, & Arana, 2011), communication (Castañer, Camerino, Anguera, & Jonsson, 2016), nutrition (Pesch & Lumeng, 2017), or clinical fields (Ruiz-Sancho, Froján-Parga, & Galván-Domínguez, 2015), among others.

Different guides are available for reporting studies based on observational methodology (Chacón-Moscoso et al., 2018; Portell, Anguera, Chacón-Moscoso, & Sanduvete-Chaves, 2015). Additionally, comprehensive guides regarding the quality of observational studies (i.e., those that examine intervention effectiveness without random assignment of participants to conditions, as described previously) have been developed (Dreyer, Bryant, & Velentgas, 2016; Vandenbroucke et al., 2014). However, no instrument has yet addressed the minimum items required to guarantee methodological quality in studies based on observational methodology (i.e., systematic and quantified recording of behavior in its natural context, without implementation of a treatment or intervention) (Anguera, 2003b). Consequently, professionals who are not experts in observational methodology do not have a basic framework to guide their practice when performing this type of study.

This work aimed to develop a checklist to measure the minimum items of methodological quality that should be considered in studies based on observational methodology. The checklist would further clarify the general guidelines presented in the Guidelines for Reporting Evaluations based on Observational Methodology (GREOM), which presents only general advice about how to conduct and report evaluations based on observational methodology (Portell et al., 2015). The present work, in contrast, intended to explicitly identify the main methodological quality items needed to conduct studies based on observational methodology, and to offer the results as a useful tool for authors performing studies and reviewers making publication decisions. Two stages were involved in developing the checklist: (a) Stage 1, validity evidence based on test content, in which we examined the items of the proposed tool, titled the Methodological Quality Checklist for studies based on Observational Methodology (MQCOM), according to the indicators of relevance (R), utility (U), and feasibility (F) (Muñiz & Fonseca-Pedrero, 2019); and (b) Stage 2, intercoder reliability (Cohen, 1960).

Stage 1: Validity Evidence Based on Test Content

Method

Participants

From among a database composed of 102 experts, 54 judges voluntarily participated. The database included academic professionals with at least 1 year of specialist experience in observational methodology and its application in sports. Recruitment was done by searching for authors of scientific articles on systematic observation in sports published from 2015 to 2018. The participants were between 25 and 75 years of age, $M = 49.5$, $SD = 12.7$, and included 33 men (61.1%) and 21 women (38.9%). They had between 1 and 40 years of experience in observational methodology, $M = 15.3$, $SD = 9.7$, and between 2 to 35 years of experience in the application of observational methodology, $M = 14.3$, $SD = 9.1$. With respect to experience, the group of participants was considered homogeneous. A statistically significant positive correlation was found between years of experience in observational methodology and years of experience in its application, $r = .719$, $p < .001$.

Instruments

An earlier study (Chacón-Moscoso et al., 2018) established the main criteria/dimensions to consider when reporting studies based on observational methodology and a list of items to measure them, based on three information sources: (1) a systematic review (Chacón-Moscoso, Sanduvete-Chaves, & Sánchez-Martín, 2016); (2) the GREOM (Portell et al., 2015), and (3) studies based on observational methodology found in 12 databases, which were of interest because of their content (Anguera & Hernández-Mendo, 2015).

This study selected and modified all items that assessed the methodological quality of studies (e.g., "Appropriateness of the instrument to the observational design") to develop the new instrument. Merely descriptive items (e.g., "Sport modality: individual or team") were omitted.

Table 1 presents the final version of the items that were presented for expert assessment to establish validity evidence based on test content, the goal being to determine which main methodological aspects should be considered in studies based on observational methodology. A total of 20 items were included that referred to 11 criteria/dimensions. Each was adjusted to successive decisions to follow the process of observational methodology: delimitation of objectives, observational design, participants/observation units, observation instrument, software use, data, specification of parameters, observational sampling, data quality control, data analysis, and interpretation of results (Anguera, Blanco-Villaseñor, & Losada, 2001; Anguera, Blanco-Villaseñor, Losada, & Portell, 2018; Chacón-Moscoso et al., 2018; Portell et al., 2015). Some criteria/dimensions consisted of only a single item because they were considered necessary to highlight the difference between items belonging to separate conceptual/methodological domains.

To examine the validity of the tool based on test content, each item was assessed by the experts on a 5-point Likert scale (Chacón-Moscoso et al., 2018; Sanduvete-Chaves, Chacón-Moscoso, Sánchez-Martín, & Pérez-Gil, 2013) in terms of three aspects with respect to its criterion/dimension (Martínez-Arias, Hernández, & Hernández, 2006). R assessed the extent to which each item was important or highlighted something regarding its criterion/dimension; non-relevant items should not be included in an instrument (Messick, 1994). U assessed how useful each item was for evaluating its assigned criterion/dimension. This aspect relates to consequential validity; an appropriate instrument must be useful to obtain the respective aim (Messick, 1989). Finally, F was understood as the possibility of recording information about each item, based on both situational factors (e.g., global attrition would be non-applicable to idiographic studies) and availability of information (information is usually given). This aspect is considered crucial in program evaluation (Chacón-Moscoso,

Table 1

Methodological Quality Checklist for studies based on Observational Methodology (MQCOM) and the Osterlind indexes obtained in Stage 1, Validity evidence based on test content

ITEMS	R	U	F
Criterion/dimension 1. Delimitation of objectives			
Item 1. Reference to observational methodology, specifying whether observation is direct or indirect: 0) Methodology is not referenced; .5) Yes, justified but not documented; 1) Yes, justified and documented.	.81	.74	.80
Item 2. Delimitation of study objectives: 0) No objectives are defined; .5) Objectives are defined behaviorally, situationally, or temporally (at least one of the three referents is missing); 1) Objectives are defined behaviorally, situationally, and temporally.	.89	.86	.87
Item 3. Theoretical framework referenced: 0) No; .33) Yes, without operationally defined constructs; .67) Yes, with partially operationally defined constructs; 1) Yes, with completely operationally defined constructs.	.75	.66	.66
Criterion/dimension 2. Observational design (Portell, Anguera, Chacón-Moscoso, & Sanduvete-Chaves, 2015)			
Item 4. Observation unit criteria (idiographic: study units are formed by one or more participants if there is a stable link between them; nomothetic: two or more study units): 0) Not identified; .5) Yes, observation units are identified, but without justification; 1) Yes, observation units are identified, with justification for the choice of an idiographic or nomothetic approach in accordance with the study objectives.	.81	.80	.80
Item 5. Temporal criteria (punctual: one or two observation sessions; follow-up: more than two observation sessions): 0) Not identified; .5) Criterion of temporality identified, but without differentiating; 1) Temporality criterion identified, differentiating between-session and within-session follow-up.	.83	.83	.79
Item 6. Dimensionality criteria (one-dimensional: one level of response; multidimensional: two or more levels of response): 0) Not identified; .5) Dimensions identified without reference to any conceptual framework; 1) Dimensions identified based on a conceptual framework.	.82	.78	.76
Criterion/dimension 3. Participants/observation units			
Item 7. Clear specification of inclusion and exclusion criteria for observation units (reasons why some units were chosen in the study and others were not): 0) No criteria or inadequate selection criteria for units, and/or with exceptions when applied; .5) Incidental (convenience, mainly accessibility); 1) Clear selection criteria for appropriate units, applied to all potential units (e.g., intentional or purposive sampling, selection based on inclusion criteria clearly specified according to the study objectives, random sampling).	.69	.71	.68
Item 8. Global attrition of units (attrition among chosen units from beginning to end): 0) 20% or more; .5) Less than 20%; 1) No; 9) Not applicable (idiographic).	**.30**	**.28**	**.31**
Criterion/dimension 4. Observation instruments			
Item 9. Adequacy of the observation instrument (combination of field format with category system, field format, category system, or scale of estimation [Anguera, 2003b]): 0) Observation instrument not available (e.g., only a list of behaviors provided); .5) Observation instrument described but not justified based on the objectives and observational design; 1) Observation instrument justified according to the objectives and observational design.	.87	.88	.82
Item 10. Codification manual with definition of the categories/behaviors and specification of dimensions (in multidimensional designs): 0) Manual not available; .5) Partial information (e.g., dimensions specified, but without definition of the categories/codes of each dimension); 1) Codification manual with definition of the categories/behaviors and specification of dimensions (in multidimensional designs).	.73	.73	.64
Criterion/dimension 5. Software use (Hernández-Mendo et al., 2014)			
Item 11. Software used to register data (SDIS-GSEQ v. 4.2.1./GSEQ 5, LINCE, MATCH VISION STUDIO, Transana, other: specify), control data quality (SDIS-GSEQ v. 4.2.1./GSEQ 5, LINCE, HOISAN, GT, SAS, other: specify), and analyze data (SDIS-GSEQ, HOISAN, THEME v. 6, R, SAS, other: specify): 0) Not used; .5) Used partially, only for some of the three aspects; 1) Used to register data, control data quality, and analyze data.	.80	.85	.84
Criterion/dimension 6. Data			
Item 12. Specification of data type (I, II, III, and IV [Bakeman, 1978]) as sequential/concurrent (sequential data: behaviors that cannot overlap and belong to a single dimension; concurrent data: behaviors that can co-occur and belong to several dimensions) and event-based/time-based (event-based: the primary parameter used in the record is order of events; time-based: the primary parameter is duration): 0) Data type not specified; .5) Data type specified but not justified; 1) Data type specified with justification; 9) Not applicable (non-sequential).	.68	.64	.64
Criterion/dimension 7. Specification of parameters			
Item 13. Type of parameters according to given use (Bakeman, 1978; Losada, 2000): 0) Primary, or basic, registration of a single category: frequency, order, and duration; .5) Secondary, derived from a single category record (ratios between primary indicators): average frequency, relative frequency, rate, relative duration, average duration, and other: specify; 1) Mixes, dynamic, or transition (two categories considered to analyze the transition from one category to another): transition frequency, relative frequency of transition, and relative duration of transition.	.57	.56	.55
Criterion/dimension 8. Observational sampling			
Item 14. Observation period (beginning and end of the segment of the behavioral flow to be studied) initially established: 0) No; .5) Unclear; 1) Yes.	**.44**	**.47**	**.44**
Item 15. Delimitation of sessions: clear establishment of criteria (temporal, behavioral, or mixed) for the beginning and the end of sessions within the observation period and of criteria for acceptance of sessions: between-sessions constancy, within-sessions constancy, or temporary disruptions (Anguera, 2003a): 0) Criteria not specified; .5) Criteria only partially specified; 1) All criteria to delimit sessions are specified.	.59	.58	.64
Item 16. Sampling rules: within-session sample of participants (Altmann, 1974): 0) No rules: ad libitum sampling; .5) Unclear; 1) Specified rules (focal: one unit of observation; scanning: more than one unit of observation, e.g. contiguous focal sampling, sequential focal sampling, alternate focal sampling; event sampling: observation units recorded if they manifest a certain event).	**.38**	**.34**	**.42**
Item 17. Within-session registration rules (Losada, 2000): 0) Not specified; .5) Unclear; 1) RAT (records activated by transitions), AO (records only of all occurrences of a category), S (sequential, records only of a sequence of categories); RAUT (records activated by units of time: intervals); instantaneous, momentary, or punctual (the last behavior that occurs at the end of the interval will be recorded); RAUT partial interval sampling (records of all categories that occur within the interval regardless of their frequency or duration); RAUT total interval sampling (records of the category present during the entire interval).	**.25**	**.25**	**.29**
Criterion/dimension 9. Data quality control			
Item 18. Between-observer reliability (agreement between the records of different observers)/within-observer reliability (agreement between the records of the same observer at two time points): 0) Not assessed; .5) Consensual agreement (qualitative); 1) Agreement is global (based on primary indicators, frequency, and duration) sequential (based on sequential-order indicators: Pearson correlation, Berk intra-class coefficient, etc.); or point-by-point (each record that each observer registers is compared): e.g., total percentage of agreement, kappa coefficient, generalizability theory).	.94	.94	.91
Criterion/dimension 10. Data analysis			
Item 19. Type of data analysis performed (Blanco-Villaseñor, Losada, & Anguera, 2003): 0) No data analysis; .33) Qualitative analysis only; .67) Descriptive analysis only; 1) Inferential analysis: relationship between categorical data (comparison of proportions); analysis of regularities (sequential analysis of delays, Markov chains, T-pattern detection, analysis of polar coordinate); multivariate analysis (logistic regression, log-linear, logit-probit, correspondence analysis); analysis of the temporal dimension (panel studies, trend analysis, time series); nonparametric tests; tests of relation (ordinal correlation, linear correlation, multiple correlation); multidimensional scaling; other: specify.	.89	.91	.84
Criterion/dimension 11. Interpretation of results			
Item 20. In the discussion section: 0) No interpretation of results or presentation of weaknesses or future developments; .5) Results are interpreted based on only the study objectives or only on cited studies in the scientific literature; and/or weaknesses or future developments are not presented; 1) Results are interpreted based on the study objectives and cited studies in the scientific literature; weaknesses and future developments are presented.	.88	.84	.82

Note: R = relevance; U = utility; F = feasibility. An item is considered appropriate when the values obtained in the three aspects measured (R, U, and F) are at least .5. Bold text indicates inappropriate results. Items 8, 14, 16, and 17 were removed from the final version of the MQCOM.

Sanduvete-Chaves, Portell, & Anguera, 2013). An open-format item was also available to allow the experts to make suggestions, such as improving the writing of an item or changing it to another more appropriate item.

The instructions were elaborated following the principles of clarity and simplicity in vocabulary (Downing & Haladyna, 2006). Participants were asked to assess the degree of R, U, and F (definitions of these three aspects were provided) for each item on a scale of 1 to 5, and to (optionally) answer the open-format item.

This instrument, created in Spanish, was translated into English following a back-translation method (International Test Commission, 2005): three experts prepared the translation, and a fourth re-translated this final version back into Spanish. Discrepancies were discussed and corrections were made.

The instrument was made available on the Internet through Google Drive Forms. Data analysis was performed using Microsoft Excel.

Procedure

Assignment of items to criteria/dimensions

All co-authors of this paper participated in ten multi-videoconference group brainstorming sessions over the course of 6 months, each lasting 2 hours and moderated by Author 1. During the sessions, the co-authors revised the items and confirmed their degree of agreement with their assigned criteria/dimensions.

Two independent coders, Author 1 and Author 5, separately assigned the 20 chosen items to the 11 criteria/dimensions, and intercoder reliability was examined by calculating Cohen's κ (Cohen, 1960). Any disagreement was resolved by consensus.

Process to obtain validity evidence based on test content

The developed instrument was sent to experts in either English or Spanish, depending on their native language, through an e-mail link to access to the instrument in Google Forms. After 15 days, we sent a reminder that the instrument was available using the same link. After another 15 days, we made the last call for answers in the same manner. After a final 15 days, the form was definitively closed to responses.

Data analysis

After gathering information, the Osterlind (1998) index of congruence was calculated for each item and each aspect (R, U, and F) was measured. The following formula was used:

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} + N\sum_{j=1}^{n} X_{ijk} - \sum_{j=1}^{n} X_{ijk}}{2(N-1)n}$$

where N = number of criteria/dimensions (11 in this case); X_{ijk} = score given by each expert to each item referring to each aspect (-1 = strongly disagree; -.5 = disagree; 0 = neither agree nor disagree; .5 = agree; and 1 = strongly agree); and n = number of experts.

Resulting scores ranged from -1 to 1. Items that obtained a score of .5 or greater on the three aspects measured were included in the final version of the MQCOM.

Results

The independent assignment of the 20 selected items to the 11 criteria/dimensions by Author 1 and Author 5 obtained an appropriate degree of inter-coder reliability, $\kappa = 1$, $p < .001$, 99% CI [1, 1].

A total of 102 experts were invited by e-mail to complete the content validity questionnaire, 54 of whom responded through Google Forms. Fifty-three of the participants sent in responses after the first call for answers, one more participant responded after the second call, and no further responses were received after the final call for answers. Table 1 shows the Osterlind indexes obtained for each item for R, U, and F. There were no statistically significant differences in participants' answers depending on their experience in observational methodology or experience in its application.

A total of 16 items achieved an Osterlind index of .5 or higher. The following four items were removed from the original 20: item 8 (global attrition of units), item 14 (observation period), item 16 (sampling rules), and item 17 (within-session registration rules).

We analyzed all the items as a whole, considering that the range of possible results was from -1 to 1. For R, *Mdn* was .77, *semi-interquartile range (SIQR)* was .142, and values ranged from .25 to .94. For U, *Mdn* was .73, *SIQR* was .142, and values ranged from .25 to .94. Finally, for F, *Mdn* was .72, *SIQR* was .124, and values ranged from .29 to .91.

Table 2 shows the open-format comments made by the experts and the actions taken to address their suggestions.

All comments were provided by one expert, except those related to item 8, which were made by two experts, and those related item 11, which were made by four experts. All experts' suggestions were considered in the final version of the proposed checklist. The suggestions concerned providing more details and specification about the content of the items, improving the definitions, and including references. Table 1 shows the final version of the MQCOM after making changes based on the results from the Osterlind indexes and the experts' comments. The final version included 16 items.

Stage 2: Intercoder reliability

To obtain indirect evidence of the instrument's validity (clear operational definitions of constructs involved in each of the instrument criteria), we conducted a study of intercoder reliability.

Method

Participants

Three co-authors of this paper participated. Author 2 (an expert in observational methodology) and Author 5 (a non-expert) coded the studies. Author 1 resolved discrepancies between the coders with respect to the meaning of the items. Each coder had a high level of written English comprehension.

Instruments

The 16-item checklist developed in Stage 1 was applied. Using simple randomization, we selected 19 papers from an updated

Table 2
Open-format comments provided by experts and resulting actions

Item[a]	Comment	Action[b]
Item 8. Global attrition of units	Regarding attrition, I understand that if it occurs in studies of direct systematic observation in real situations, it has a minor importance (with another meaning) compared in selective studies with longitudinal design.	This item was eliminated because it did not fulfill the inclusion criteria in the study of validity based on test content.
Item 9. Adequacy of the observation instrument	One might think that item 9 could raise more doubt for non-experts in terms of definition.	The description of the item was simplified, and a reference was added.
Item 10. Codification manual	I would explain with an example of each the categories.	An example of each category was added.
Item 11. Software use	Missing references for papers regarding LINCE and HOISAN.	Missing references were included.
Item 19. Data analysis	You could include more items related to the type of data analysis that this methodology allows, e.g., descriptive, inferential, regression, and sequential.	Partially done. Different types of analysis were included and delimited in greater detail within the same item (additional items were not added).

Note: Only items that received comments have been included in Table 2
[a] Items appear in abbreviated form; full items are available in Table 1. [b] The changes resulting from the experts' comments are reflected in the final version of the checklist (Table 1)

computerized database of studies that applied observational methodology to investigate soccer (see Appendix). SPSS 25.0 was used to calculate Cohen's κ.

Procedure

The two coders (Author 2 and Author 5) were trained in the application of the checklist. First, they discussed each item and the options. Author 1 mediated when discrepancies were difficult to resolve. Next, both coders independently applied the checklist to a single paper based on observational methodology. Before the two coders independently coded the selected papers, any discrepancies were resolved with the arbitration of Author 1. Once the coders completed the training, 19 numbers were randomly generated without repetition using a website. The 19 papers corresponding to the randomly generated numbers were selected from the database of numbered studies already collected. The two coders independently applied the checklist developed in Stage 1 to the selected papers.

Data analysis

Cohen's κ was calculated for each item to study the concordance between coders. Values over .7 were considered to show adequate intercoder reliability (López-Pina et al., 2015).

Results

Table 3 presents the intercoder reliability results obtained for each item. All 16 items obtained statistically significant κ values above .7. Three items (6, 10, and 20) obtained an agreement value between .75 and .8; five (items 3, 7, 12, 15, and 19) obtained an agreement value between .8 and .9; and eight (items 1, 2, 4, 5, 9, 11, 13, and 18) obtained an agreement value between .9 and 1. Four of these (items 1, 4, 11, and 18) obtained the highest possible agreement.

The CI provided information regarding the accuracy of results and ranged in amplitude from 0, 99% CI [1 – 1] (items 1, 4, 11, and 18) to .607, 99% CI [.393 – 1] (items 6 and 10).

Discussion

This study proposed a simple, relevant, and useful 16-item checklist designed for use by intervention professionals applying observational methodology in various areas. The checklist, called the MQCOM, includes individual methodological features that serve as quality indicators to be considered when designing, implementing, or evaluating a study based on observational methodology.

The study further proposed the necessary main criteria/dimensions to consider in observational methodology, specifically concerning successive methodological decisions to follow the process of observational methodology, and explicitly clarified the minimum operational items that should be considered; thus, some criteria/dimensions include only a single item.

Table 3
Results of intercoder reliability testing

Item	κ	99% CI
1	1**	[1–1]
2	.905**	[.498–1]
3	.824**	[.439–1]
4	1**	[1–1]
5	.905**	[.498–1]
6	.755**	[.393–1]
7	.824**	[.439–1]
9	.905**	[.498–1]
10	.755**	[.393–1]
11	1**	[1–1]
12	.824**	[.439–1]
13	.905**	[.498–1]
15	.824**	[.439–1]
18	1**	[1–1]
19	.824**	[.439–1]
20	.765**	[.409–1]

Note: Missing items were removed in Stage 1. κ = Cohen's κ coefficient.
** $p < .01$

The main strengths of this paper are that it presents a new checklist answering a clear need in this applied area, it is based on an exhaustive literature review (research group results obtained over the last 30 years), and it employed assessment of quantitative and qualitative data by a significant number of judges (19 of whom had more than 15 years of experience in observational methodology). According to Prieto and Muñiz (2000), a wide number of experts were consulted (54, $N > 30$). All suggestions made by the judges to improve the checklist were considered in the final version.

The study has a possible limitation regarding the relatively small number of papers included in the intercoder reliability study and their level of specificity (observational methodology applied to soccer only). Nevertheless, the objective of the exploratory intercoder reliability study was to analyze the reliability of the checklist, not its generalizability. We consider this study a starting point from which to apply the MQCOM extensively in various research areas based on observational methodology. We invite readers and potential users to apply this checklist and share their results to further research of its dimensionality, level of invariance, or possible points of convergence or divergence with other instruments applied in different types of methodologies.

Acknowledgments

The authors greatly appreciate all comments and suggestions received from the experts, the reviewers, and the English language editor. We consider this work to be improved thanks to them.

This research was funded by the Chilean National Fund of Scientific and Technological Development, FONDECYT Regular (CONICYT, ref. no. 1190945), by the Operational Program FEDER Andalusia, Junta de Andalucía (ref. no. US-1263096), and by the Ministerio de Ciencia, Innovación y Universidades, Programa Estatal de Generación de Conocimiento y Fortalecimiento Científico y Tecnológico del Sistema I+D+i (ref. no. PGC2018-098742-B-C31) (2019-2021): as part of the coordinated project *New approach of research in physical activity and sports from a mixed methods perspective* (NARPAS_MM, ref. no. SPGC201800X098742CV0). We gratefully acknowledge the support of the Generalitat de Catalunya Research Group (GRUP DE RECERCA E INNOVACIÓ EN DISSENYS [GRID]) Tecnología i aplicació multimedia i digital als dissenys observacionals (grant no. 2017 SGR 1405).

References

Altmann, J. (1974). Observational study of behaviour: Sampling methods. *Behaviour, 49*, 227-267.

Anguera, M. T. (1979). Observational typology. *Quality & Quantity, 13*, 449-484. doi:10.1007/BF00222999

Anguera, M. T. (1996). Introduction. Monograph on observation in assessment. *European Journal of Psychological Assessment*, *12*, 87-88.

Anguera, M. T. (2003a). La observación [The observation]. In C. Moreno Rosset (Ed.), *Evaluación psicológica. Concepto, proceso y aplicación en las áreas del desarrollo y de la inteligencia* (pp. 271-308). Madrid, Spain: Sanz y Torres.

Anguera, M. T. (2003b). Observational Methods (General). In R. Fernández-Ballesteros (Ed.), *Encyclopedia of Psychological Assessment, Vol. 2* (pp. 632-637). London, UK: Sage.

Anguera, M. T., Blanco-Villaseñor, A., & Losada, J. L. (2001). Diseños observacionales, cuestión clave en el proceso de la metodología observacional [Observational designs, a key issue in the process of observational methodology]. *Metodología de las Ciencias del Comportamiento, 3*(2), 135-160.

Anguera, M. T., Blanco-Villaseñor, A., Losada, J. L., & Portell, M. (2018). Pautas para elaborar trabajos que utilizan la metodología observacional [Guidelines for designing and conducting a study that applies observational methodology]. *Anuario de Psicología, 48*(1), 9-17. doi:10.1016/j.anpsic.2018.02.001

Anguera, M. T., & Hernández-Mendo, A. (2015). Técnicas de análisis en estudios observacionales en ciencias del deporte [Analyses techniques studies in sport science]. *Cuadernos de Psicología del Deporte, 15*(1), 13-30.

Bakeman, R. (1978). Untangling streams of behavior: Sequential analysis of observation data. In G. P. Sackett (Ed.), *Observing Behavior: Vol. 2. Data collection and analysis methods* (pp. 63-78). Baltimore, MD: University of Park Press.

Blanco-Villaseñor, A., Losada, J. L., & Anguera, M. T. (2003). Data analysis techniques in observational designs applied to the environment-behaviour relation. *Medio Ambiente y Comportamiento Humano, 4*(2), 111-126.

Castañer, M., Camerino, O., Anguera, M. T., & Jonsson, G. K. (2016). Paraverbal communicative teaching T-patterns using SOCIN and SOPROX observational systems. In J. K. Burgoon, M. Magnusson, & M. Casarrubea (Eds.), *Discovering hidden temporal patterns in behavior and interaction* (pp. 83-100). New York, NY: Springer.

Chacón-Moscoso, S., Sanduvete-Chaves, S., Anguera, M. T., Losada, J. L., Portell, M., & Lozano-Lozano, J. A. (2018). Preliminary checklist for reporting observational studies in sports areas: Content validity. *Frontiers in Psychology*, *9*, 291. doi:10.3389/fpsyg.2018.00291

Chacón-Moscoso, S., Sanduvete-Chaves, S., Portell, M., & Anguera, M. T. (2013). Reporting a program evaluation: Needs, program plan, intervention, and decisions. *International Journal of Clinical and Health Psychology, 13*(1), 58-66. Retrieved from http://www.aepc.es/ijchp/articulos_pdf/ijchp-433_es.pdf

Chacón-Moscoso, S., Sanduvete-Chaves, S., & Sánchez-Martín, M. (2016). The development of a checklist to enhance methodological quality in intervention programs. *Frontiers in Psychology*, *7*, 1811. doi:10.3389/fpsyg.2016.01811

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Dreyer, N. A., Bryant, A., & Velentgas, P. (2016). The GRACE checklist: A validated assessment tool for high quality observational studies of comparative effectiveness. *Journal of Managed Care & Specialty Pharmacy*, *22*(10), 1107-1113. doi:10.18553/jmcp.2016.22.10.1107

Garzón, B., Lapresa, D., Anguera, M. T. & Arana, J. (2011). Análisis observacional del lanzamiento de tiro libre en jugadores de baloncesto base [Observational analysis of the free throw shot made by grassroots basketball players]. *Psicothema, 23*(4), 851-857.

Gimeno, A., Anguera, M. T., Berzosa, A., & Ramírez, L. (2006). Detección de patrones interactivos en la comunicación de familias con hijos adolescentes [Interactive patterns detection in family communication with adolescents]. *Psicothema, 18*(4), 785-790.

Hernández-Mendo, A., Castellano, J., Camerino, O., Jonsson, G., Blanco-Villaseñor, A., Lopes, A., & Anguera, M. T. (2014). Programas informáticos de registro, control de calidad del dato, y análisis de datos [Observational software, data quality control and data analysis]. *Revista de Psicología del Deporte, 23*(1), 111-121.

International Test Commission (2005). *ITC guidelines for translating and adapting tests*. Retrieved from https://www.intestcom.org/files/guideline_test_adaptation.pdf

López-Pina, J. A., Sánchez-Meca, J., López-López, J. A., Marín-Martínez, F., Núñez-Núñez, R. M., Rosa-Alcázar, A. I., … Ferrer-Requena, J. (2015).

The Yale–Brown Obsessive Compulsive Scale: A reliability generalization meta-analysis. *Assessment, 22*(5), 619-628.

Losada, J. L. (2000). *Metodología observacional* [Observational methodology]. A Coruña, Spain: Penta.

Martínez-Arias, M. R., Hernández, M. J., & Hernández, M. V. (2006). *Psicometría* [Psychometrics]. Madrid, Spain: Alianza Editorial.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-104). New York, NY: MacMillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test [Ten steps for test development]. *Psicothema, 31*(1), 7-16. doi:10.7334/psicothema2018.291

Osterlind, S. J. (1998). *Constructing tests items*. Boston, MA: Kluwer Academic Publishers.

Pesch, M. H., & Lumeng, J. C. (2017). Methodological considerations for observational coding of eating and feeding behaviors in children and their families. *International Journal of Behavioral Nutrition and Physical Activity*, *14*(1), 170. doi:10.1186/s12966-017-0619-3

Portell, M., Anguera, M. T., Chacón-Moscoso, S., & Sanduvete-Chaves, S. (2015). Guidelines for Reporting Evaluations based on Observational Methodology (GREOM). *Psicothema, 27*(3), 283-289. doi:10.7334/psicothema2014.276

Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los test utilizados en España [A model to evaluate the quality of tests used in Spain]. *Papeles del Psicólogo, 77*, 65-75.

Ruiz-Sancho, E., Froján-Parga, M. X., & Galván-Domínguez, N. (2015). Verbal interaction patterns in the clinical context: A model of how people change in therapy. *Psicothema, 27*(2), 99-107. doi:10.7334/psicothema2014.119

Sanduvete-Chaves, S., Chacón-Moscoso, S., Sánchez-Martín, M., & Pérez-Gil, J. A. (2013). The revised Osterlind index: A comparative analysis in content validity studies. *Acción Psicológica*, *10*(2), 10-26. doi:10.5944/ap.10.2.11821

Santoyo, C., Jonsson, G. K., Anguera, M. T., & López-López, J. A. (2017). Observational analysis of the organization of on-task behavior in the classroom using complementary data analyses. *Anales de Psicología*, *33*(3), 497-514. doi:10.6018/analesps.33.3.271061

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., & STROBE Initiative (2014). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *International Journal of Surgery, 12*(12), 1500-1524.

Appendix
List of papers selected and used in stage 2 to test intercoder reliability

Amatria, M., Lapresa, D., Arana, J., Anguera, M. T., & Garzón, B. (2016). Optimization of game formats in U-10 soccer using logistic regression analysis. *Journal of Human Kinetics, 54*, 163-171. doi:10.1515/hukin-2016-0047

Amatria, M., Lapresa, D., Arana, J., Anguera, M. T., & Jonsson, G. K. (2017). Detection and selection of behavioral patterns using Theme: A concrete example in grassroots soccer. *Sports, 13*(5), E20. doi:10.3390/sports5010020.

Arana, J., Lapresa, D., Anguera, M. T., & Garzón, B. (2013). Adapting football to the child: An application of the logistic regression model in observational methodology. *Quality & Quantity*, *47*(6), 3473-3480. doi:10.1007/s11135-012-9734-z

Barreira, D., Garganta, J., Castellano, J., Machado, J., & Anguera, M. T. (2015). How elite-level soccer dynamics has evolved over the last three decades? Input from generalizability theory. *Cuadernos de Psicología del Deporte*, *15*(1), 51-62.

Barreira, D., Garganta, J., Castellano, J., Prudente, J., & Anguera, M. T. (2014). Evolución del ataque en el fútbol de élite entre 1982 y 2010: Aplicación del análisis secuencial de retardos [Evolution of attacking patterns in elite-level soccer between 1982 and 2010: The application of lag sequential analysis]. *Revista de Psicología del Deporte*, *23*(1), 139-146.

Barreira, D., Garganta, J., Guimarães, P., Machado, J., & Anguera, M. T. (2014). Ball recovery patterns as a performance indicator in elite soccer. *Journal of Sports Engineering and Technology, 228*(1), 61-72. doi:10.1177/1754337113493083

Barreira, D., Garganta, J., Machado, J. C., & Anguera, M. T. (2014). Effects of ball recovery in top-level soccer attacking patterns of play. *Revista Brasileira de Cineantropometria & Desempenho Humano, 16*(1), 36-46. doi:10.5007/1980-0037.2014v16n1p36

Casal, C. A., Andujar, M. A., Losada, J. L., Ardá, T., & Maneiro, R. (2016). Identification of defensive performance factors in the 2010 FIFA World Cup South Africa. *Sports, 4*, 54. doi:10.3390/sports4040054

Casal, C. A., Losada, J. L., & Ardá, A. (2015). Análisis de los factores de rendimiento de las transiciones ofensivas en el fútbol de alto nivel [Analysis of the performance factors of the offensive transitions in high level football]. *Revista de Psicología del Deporte*, *24*(1), 103-110.

Casal, C. A., Maneiro, R., Ardá, T., Losada, J. L., & Rial, A. (2014). Effectiveness of indirect free kicks in elite soccer. *International Journal of Performance Analysis in Sport*, *14*(3), 744-760. doi:10.1080/24748668.2014.11868755

Casal, C. A, Maneiro, R., Ardá, T., Marí, F. J., & Losada, J. L. (2017). Possession zone as a performance indicator in football. The game of the best teams. *Frontiers in Psychology*, 8, 1176. doi:10.3389/fpsyg.2017.01176

Castellano, J., & Blanco-Villaseñor, A. (2015). Análisis de la variabilidad del desplazamiento de futbolistas de élite durante una temporada competitiva a partir de un modelo lineal mixto generalizado [Analysis of the variability of the movement of elite soccer players during a competitive season of a generalized linear mixed model]. *Cuadernos de Psicología del Deporte, 15*(1), 161-168.

Echeazarra, I., Castellano, J., Usabiaga, O., & Hernández-Mendo, A. (2015). Comparación del uso del espacio en categorías infantil y cadete de fútbol a partir del análisis de coordenadas polares [Strategic use of space in under 14 and under 16 soccer: A polar coordinate analysis]. *Cuadernos de Psicología del Deporte*, *15*(1), 169-180.

Lapresa, D., Arana, J., Amatria, M., Fernández, F. J., & Anguera, M. T. (2017). Fútbol: efectos de una unidad didáctica en la iniciación temprana [Soccer: Effects of a teaching unit in early initiation]. *Apunts. Educación Física y Deportes, 1*(127), 59-67. doi:10.5672/apunts.2014-0983.es.(2017/1).127.06

Lapresa, D., Arana, J., Anguera, M. T., Pérez-Castellanos, J. I., & Amatria, M. (2016). Application of logistic regression models in observational methodology: Game formats in grassroots football in initiation into football. *Anales de Psicología*, *32*(1), 288-294. doi:10.6018/analesps.31.3.186951

Lapresa, D., Arana, J., & Garzón, B. (2006). El fútbol 9 como alternativa al fútbol 11, a partir del estudio de la utilización del espacio de juego [9 football as an adjustment alternative to 11 football, based on the control of the space]. *Apunts. Educación Física y Deportes, 4*(86), 34-44.

Lapresa, D., Arana, J., Garzón, B., Egüen, R., & Amatria, M. (2010). Adaptando la competición en la iniciación al fútbol: estudio comparativo de las modalidades de fútbol 3 y fútbol 5 en categoría prebenjamín [Adapting competition in beginners' football: A comparative study of 3-a-side football and 5-a-side football in the under-eights. *Apunts. Educación Física y Deportes, 3*(101), 43-56.

Lapresa, D., Arana, J., Ugarte, J., & Garzón, B. (2009). Análisis comparativo de la acción ofensiva en F-7 y F-8, en la categoría alevín [Contrastive analysis of the offensive part in 7-8- football game, in 12 years old age]. *Retos: Nuevas Tendencias en Educación Física, Deporte y Recreación, 16*, 97-103.

Maneiro, R., Losada, J. L., Casal, C. A., & Ardá, A. (2017). Análisis multivariante del tiro libre indirecto en la Copa del Mundo de la FIFA 2014 [Multivariate analysis of indirect free kick in the FIFA World Cup 2014]. *Anales de Psicología*, *33*(3), 461-470.